МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И
КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра системного анализа и компьютерного моделирования

СИКОЛЕНКО Максим Александрович

СОЗДАНИЕ ПРОГРАММНОГО КОНВЕЙЕРА ДЛЯ ПОЛУЧЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ ЗНАЧИМОЙ ЧАСТИ ГЕНОМА ВИРУСА SARS-COV-2 ИЗ ПРОЧТЕНИЙ ПЕРЕКРЫВАЮЩИХСЯ АМПЛИКОНОВ

Аннотация к магистерская диссертация

специальность 1-98 80 01 «Информационная безопасность»

Научный руководитель
Леонид Николаевич Валентович
кандидат биологических наук, доцент

Допущена к защите
«___» _____ 2022 г.
Зав. кафедрой системного анализа
и компьютерного моделирования
_____В.В. Скакун

Минск, 2022

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Ключевые слова: SARS-CoV-2, геном, секвенирование, программный конвейер, предобработка, картирование, аннотация, праймеры, Illumina, Oxford Nanopore Technologies.

Определение нуклеотидной последовательности геномов возбудителей инфекционных заболеваний способствует снижению неопределённости в отношении распространения, разнообразия и изменчивости патогенов. Вирус SARS-CoV-2, возбудитель опасного инфекционного заболевания COVID-19, является важным объектом изучения эпидемиологов из-за продолжающейся пандемии COVID-19.

Объект исследования: программное обеспечение, определяющее последовательности полных геномов штаммов вируса SARS-CoV-2.

Предмет исследования: максимизация достоверности результирующих геномных последовательностей.

Необходимым этапом определения последовательности генома является компьютерная обработка нуклеотидных последовательностей (т .н. прочтений). Особенности процесса секвенирования генома SARS-CoV-2 таковы, что для обработки прочтений необходимо, помимо универсально применяемых методов обработки нуклеотидных данных, применять специальные методы, учитывающие особенности пробоподготовки. Существующие программы допускают ряд недочётов при обработке таких данных; эти недочёты могут приводить к искажениям в результирующей последовательности. Таким образом, цель данной работы заключалась в том, чтобы разработать программный конвейер, предназначенный для определения последовательности значимой части генома вируса SARS-CoV-2, сводящий к минимуму ошибки в результирующей последовательности.

В результате был разработан ряд компьютерных программ, реализующий полный процесс получения последовательности генома вируса SARS-CoV-2: от исходных прочтений до проаннотированной результирующей последовательности. При этом конвейер использует метод предобработки прочтений, минимизирующий искажения в результирующей последовательности, превосходящий аналогичные подходы.

Объём работы: 75 страниц. 17 иллюстраций, 1 таблица, 1 приложение. 60 использованных библиографических источников.

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Ключавыя словы: SARS-CoV-2, геном, секвенаванне, праграмны канвеер, перадапрацоўка, картаванне, анатацыя, праймеры, Illumina, Oxford Nanopore Technologies.

Вызначэнне нуклеатыднай паслядоўнасці геномаў узбуджальнікаў інфекцыйных захворванняў спрыяе паніжэнню нявызначанасці наконт распаўсюду, разнастайнасці патагена і таго, як ён эвалюцыянуе. Вірус SARS-CoV-2, узбуджальнік небяспечнага інфекцыйнага захворвання COVID-19, з'яўляецца важным аб'ектам даследаванняў з боку эпідэміёлагаў праз бягучую пандэмію COVID-19

.

Аб'ект даследавання: праграмнае забеспячэнне, якое вызначае паслядоўнасці поўных геномаў штамаў вірусаў SARS-CoV-2.

Прадмет даследавання: максімізацыя дакладнасці выніковых геномных паслядоўнасцяў.

Неабходным этапам вызначэння паслядоўнасці генома з'яўляецца камп'ютарная апрацоўка нуклеатыдных паслядоўнасцяў (г. зв. прачытанняў). Асаблівасці працэсу секвенавання генома SARS-CoV-2 такія, што для апрацоўкі прачытанняў неабходна выкарыстоўваць — апроч тых метадаў апрацоўкі нуклеатыдных дадзеных, якія выкарыстоўваюцца ўніверсальна — спецыяльныя метады, якія ўлічваюць асаблівасці падрыхтоўкі матэрыялу для секвенавання. Існыя праграмы дапускаюць шэраг хібаў пры апрацоўцы такіх дадзеных; такія хібы могуць прыводзіць да скажэнняў у выніковай паслядоўнасці. Такім чынам, мэта гэтай работы зводзіцца да таго, каб распрацаваць праграмны канвеер, які б ажыццяўляў вызначэнне паслядоўнасці значнай часткі генома віруса SARS-CoV-2 і пры гэтым мінімізаваў бы колькасць памылак у выніковай паслядоўнасці.

У выніку быў распрацаваны шэраг камп'ютарных праграм, які рэалізуе працэс вызначэння паслядоўнасці генома віруса SARS-CoV-2: ад зыходных прачытанняў да выніковай паслядоўнасці з анатацыяй. Пры гэтым канвеер выкарыстоўвае спосаб перадапрацоўкі прачытанняў, які мінімізуе скажэнні ў выніковай паслядоўнасці і пераўзыходзіць аналагічныя падыходы.

Аб'ём работы: 75 старонак. 17 ілюстрацый, 1 табліца, 1 дадатак. 60 выкарыстаных бібліяграфічных крыніц.

RESUME

Key words: SARS-CoV-2, genome, sequencing, pipeline, preprocessing, mapping, annotation, primers, Illumina, Oxford Nanopore Technologies.

Determination of genomic sequences of pathogens helps to minimize the uncertainty about spread, diversity and variability of pathogens. The SARS-CoV-2 virus, which causes dangerous infectious disease COVID-19, is an important object of study for epidemiologists due to the ongoing COVID-19 pandemics.

Subject of the study: software, which determines whole-genome sequences of SARS-CoV-2 virus strains.

Topic of the study: maximization of accuracy of resultant genomic sequences.

Computer processing of nucleotide sequences obtained with devices called sequencers (such sequences are called reads) is an essential step of a genome sequence determination. The specifics of SARS-CoV-2 genome sequencing require use of specialized computer programs which take into account the specific features of sample preparation, along with programs for common and universal nucleotide data processing. The existing programs are prone to errors, which may lead to resulting genome sequence being erroneous. Therefore, the aim of the current project was to develop a pipeline for determination of sequence of the significant part of SARS-CoV-2 genome which minimizes the amount of errors in the resulting sequence.

As a result of the current project, a pipeline was developed implementing the process of determining of sequence of SARS-CoV-2 genome — from "raw" reads to annotated result sequence. The pipeline uses a method of reads preprocessing, which minimizes errors introduced to resulting sequence and permits less errors than analogous methods.

Size: 75 pages, 17 figures, 1 table, 1 supplementary material. 60 used bibliographical sources.